BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 1 2008, pages 78–85
doi:10.1093/bioinformatics/btm531

Gene expression

GlobalANCOVA: exploration and assessment of gene group effects

Manuela Hummel[1,*], Reinhard Meister[2] and Ulrich Mansmann[1,3]

[1]IBE, University of Munich, [2]Technische Fachhochschule Berlin and [3]Department of Statistics, University of Munich, Germany

Received on July 9, 2007; revised on September 24, 2007; accepted on October 16, 2007

Advance Access publication November 17, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Several authors have studied expression in gene sets with specific goals: overrepresentation of interesting genes in functional groups, predictive power for class membership and searches for groups where the constituent genes show coordinated changes in expression under the experimental conditions. The purpose of this article is to follow the third direction. One important aspect is that the gene sets under analysis are known a priori and are not determined from the experimental data at hand. Our goal is to provide a methodology that helps to identify the relevant structural constituents (phenotypical, experimental design, biological component) that determine gene expression in a group.

Results: Gene-wise linear models are used to formalize the structural aspects of a study. The full model is contrasted with a reduced model that lacks the relevant design component. A comparison with respect to goodness of fit is made and quantified. An asymptotic test and a permutation test are derived to test the null hypothesis that the reduced model sufficiently explains the observed expression within the gene group of interest. Graphical tools are available to illustrate and interpret the results of the analysis. Examples demonstrate the wide range of application.

Availability: The R-package `GlobalAncova` (http://www.bioconductor.org) offers data and functions as well as a vignette to guide the user through specific analysis steps.

Contact: hummel@ibe.med.uni-muenchen.de

1 INTRODUCTION

The concept of a gene group is quite vague but often used. It emphasizes the conceptual and biological importance of the expression profile of a group of genes versus the multiple analysis of individual gene expression. So far, there are two major strategies to perform group tests: (1) prove that a group is outstanding in its expression compared to other groups or all remaining genes, (2) prove that expression in a given group is altered between different conditions of interest. Goeman and Bühlman (2007) call the null hypotheses related to (1) *competitive* and related to (2) *self-contained*. In order to assess (2), some authors apply (1) by assuming that there is no differential expression between the experimental conditions for all genes not contained in the gene group of interest. From a statistical point of view, this is not an optimal way to proceed. This article offers a strategy to handle problems that can be formulated in terms of *self-contained* null hypotheses.

A famous example for a *competitive* null hypothesis was given by Lamb *et al*. (2003). They looked at the set of target genes of the cyclin D1 gene and study the target set expression profile within different tumours. They showed that the group's overall expression level is significantly elevated compared to the expression level of a randomly chosen group of genes of the same size.

Gene signatures are specific gene groups that constitute a classifier to discriminate between conditions (e.g. phenotypes). A first validation step for a signature on new data consists in an evaluation of its predictive ability between conditions and to summarize the visual impression of a heat map in an informative statistical measure. The validation consists in testing a *self-contained* null hypothesis.

In general, the functional interpretation of a gene signature is not straightforward. The direct involvement of its single components in biological processes needs careful analysis. The signature's functional annotation may start by assessing coexpression of signature genes with genes of relevant functional groups. Again, one must consider *self-contained* null hypotheses.

Molecular studies have been largely focused on individual candidate genes, contrasting with the molecular complexity of disease. The concept of a pathway tries to grasp this complexity and to open the view for a more appropriate biological understanding. However, the dependencies and dynamics of interest are not easily formalized. Pathways can be conceptualized as gene sets enriched with structure (implied by networks and/or dynamics). Therefore, the analysis of gene groups offers a crude approach to pathway analysis. Groene *et al*. (2006) have correlated tumour stages of colorectal cancer samples with gene activity in nine well defined cancer-related pathways. They identified pathways where the transcription pattern between both stages showed a clear distinction (*self-contained* null hypotheses). But, they did not ask the *competitive* question: is there a pathway whose transcriptional differences between samples of both stages is outstanding compared to those of the other pathways?

*To whom correspondence should be addressed.

Strategies to test *competitive* null hypotheses have been developed by several authors. Typically genes are scored according to a rule. *Gene set enrichment strategies* use the observed or absolute value of the gene-wise test statistics. Strategies based on the *hypergeometric distribution* score genes by 0 or 1, respectively, if the *P*-value of the gene-wise test is above or below a fixed cutoff value. The score value distribution is compared between gene sets.

Gentleman and Falcon (2007) define a set of categories as merely a grouping of genes (entities). The groups do not need to be exhaustive or disjoint. The mapping from a set of entities (genes) to a set of categories can be represented as a bipartite graph, whereby one set of nodes are the genes and the other are the categories. The category approach answers questions such as whether the observed test statistic is unusual for a given category or whether any of the observed category statistics are unusually large or small with respect to the entire reference distribution. Therefore, the category approach can be used to study both *self-contained* and *competitive* null hypotheses.

Goeman *et al*. (2004) introduced the concept of a *global test*. A *global null hypothesis* is the aggregation of many individual null hypotheses. The global null hypothesis related to a gene group is a statement that applies to all individual genes contained in the group: *no gene in the group exhibits differential expression between the conditions of interest*. Goeman *et al*. (2004) evaluated the influence of an expression profile on a phenotype. Their approach is motivated by the validation problem for gene signatures: does the knowledge of a profile help to improve the prediction of the phenotype (group, quantitative trait, survival)? Their approach is related to prediction, and this introduces logical constraints on the situations where *global test* can be used.

Global tests for the specific situation of a group comparison without adjustment for covariates were developed by Kong *et al*. (2006). They use Hotelling's T^2 test, the multidimensional analog of the univariate two-sample *t*-test, that accounts for correlation between genes. The dimension problem in the case of having a larger number of genes in the gene set compared to the number of samples is addressed by the use of a principal component approach.

It is the purpose of this article to offer a general methodology to study *how the expression structure within a group of genes is influenced by design aspects of the study (experiment)*. Therefore, the article studies *self-contained* null hypotheses and demonstrates the need to develop more general approaches beyond the *category approach* and the *global test*. Gene-wise linear models are used to formalize the relationship of gene expression with phenotypic or genomic covariates. An ANOVA-based sum of squares summarizes the individual gene-wise linear models to a group statement. This provides the name of our procedure: GlobalANCOVA. A permutation test and an asymptotic distribution of the test statistics under the null hypothesis are available to calculate *P*-values. This work extends a former version of GlobalANCOVA (Mansmann and Meister, 2005), which was confined to two-group comparisons. It considers a broader range of designs by exploiting the full scope of linear model theory.

Linear models have been successfully used to analyse gene expression experiments. They were introduced by Kerr *et al*. (2000) to simultaneously normalize and analyse gene expression data. Smyth (2005) used linear models and ANOVA when they developed *limma*. GlobalANCOVA extends limma in two directions: First, the use of multiple comparisons between many RNA targets is replaced by a simultaneous global assessment for the entire group of genes. Second, a tool based on linear models is proposed with a wider range of applications beyond designed experiments.

The following section introduces the model in a formal way. A simulation study is performed to assess the statistical properties of GlobalANCOVA. Graphical tools are shown to visualize the results of a GlobalANCOVA analysis. Examples that cover a wide range of novel applications will be presented. The discussion offers a broader view on the potential of group testing.

2 METHODS

2.1 The basic concept

The general framework looks as follows: p genes are measured in n independent samples (not necessarily $n<<p$). Additionally d phenotypic covariates are documented for each sample. To illustrate the basic aspects of the formalism, we introduce a toy example from oncology. Table 1 shows the covariate information ($d=3$) of samples from eight patients who belong to two groups (0—good prognosis, 1—bad prognosis) with additional phenotypic information on sex and localization of the probe material.

A gene-specific linear model quantifies the systematic part, $\tilde{m}^{(i)}$ and the noise component, $\tilde{\xi}^{(i)}$ of the expression measurements for gene i over the n samples, $\tilde{x}^{(i)} = (x_1^i, \ldots, x_n^i)^t$. The model for gene i is described by the gene-specific $(d+1)$ dimensional regression coefficient $\tilde{\beta}_i$ and the design matrix C that is independent of gene i:

$$\tilde{x}^{(i)} = \tilde{m}^{(i)} + \tilde{\xi}^{(i)} = \begin{pmatrix} 1 & c_{11} & \cdots & c_{1d} \\ \vdots & \vdots & & \vdots \\ 1 & c_{n1} & \cdots & c_{nd} \end{pmatrix} \cdot \begin{pmatrix} \beta_{i0} \\ \vdots \\ \beta_{id} \end{pmatrix} + \begin{pmatrix} \xi_1^i \\ \vdots \\ \xi_n^i \end{pmatrix} = C \cdot \tilde{\beta}_i^t + \tilde{\xi}^{(i)} \quad (1)$$

The phenotypic covariate vector for sample k is $\tilde{c}_k = (c_{k1}, \ldots, c_{kd})$. The value 1 in the first column of the design matrix C defines a gene-specific mean expression that is quantified by β_{i0}. The other parts of the regression coefficient $\tilde{\beta}_i$ quantify the mean influence of the specific covariates on the expression of the gene. The noise component for gene i $\tilde{\xi}^{(i)}$ has a mean of zero and an unspecified diagonal covariance matrix $\mathrm{Cov}[\tilde{\xi}^{(i)}] = \sigma_i^2 \cdot \Delta_n$, where Δ_n is the n dimensional unity diagonal matrix.

The toy example helps to illustrate the formalism that the four components of $\tilde{\beta}_i$ have the following interpretation: the first component quantifies the overall gene-specific mean expression adjusted for all

Table 1. Design matrix for a simple two-group setting with adjustment for sex (1—male; 0—female) and location (1—colon; 0—rectum)

Samples	S1	S2	S3	S4	S5	S6	S7	S8
Gene (*i*) specific mean	1	1	1	1	1	1	1	1
Group	0	0	0	0	1	1	1	1
Sex	1	1	0	0	0	0	1	1
Localization	1	0	1	0	1	0	1	0

covariate effects. The second component quantifies the corrected (for sex and location effects) mean differential gene expression between patients with good and bad prognoses. The third and fourth components describe mean differences in gene expression between male and female patients and mean differences in gene expression between samples taken from rectum or colon. The design matrix C is given by transposing the rows of Table 1 to columns.

The aim of GlobalANCOVA is to prove the relevance of certain covariates in explaining the observed gene expression, called covariates of interest. Therefore, two models are compared: the full model, which contains all covariates (FM) and the reduced model (RM), which does not have the covariates of interest. Formally, the design matrix C and the regression coefficient $\tilde{\beta}_i$ are divided into the corresponding parts. The submatrix C_0 contains the covariates of interest, the submatrix C_1 contains the remaining covariates used for adjustment. For gene i, the prediction of expression under the full or reduced model are:

$$\begin{aligned} E(\tilde{x}^{(i)}_{\mathrm{FM}}) &= [C_0, C_1](\tilde{\beta}_{i,0}, \tilde{\beta}_{i,1})^t \\ \text{and} \quad E(\tilde{x}^{(i)}_{\mathrm{RM}}) &= C_1 \tilde{\beta}^t_{i,1} \end{aligned} \tag{2}$$

For the toy example the matrix C_0 is simply the column vector containing the group information, and C_1 is the 8×3 matrix with the columns defined by the constant vector for the gene-specific mean, the information on sex and the localization.

The residual sum of squares (RSS) quantifies how well a prediction fits the observed data and quantifies the ability of a model to explain the observed data. It is necessary to assemble the single gene information in a global linear model. Model comparison in linear model theory proceeds by defining suitable measures to compare two residual sums of squares and to offer appropriate statistical tests for the null hypothesis that *both models explain the data equally well* (Draper and Smith 1998). The relevance of certain covariates in explaining the observed gene expression is proven if the full model explains the observation better than the reduced model.

The gene-wise information is assembled in a global linear model:

$$\tilde{X} = \begin{pmatrix} \tilde{x}^{(1)} \\ \vdots \\ \tilde{x}^{(p)} \end{pmatrix} = \begin{pmatrix} C & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & C \end{pmatrix} \cdot \begin{pmatrix} \tilde{\beta}^t_1 \\ \vdots \\ \tilde{\beta}^t_p \end{pmatrix} + \begin{pmatrix} \tilde{\xi}^{(1)} \\ \vdots \\ \tilde{\xi}^{(p)} \end{pmatrix} = \tilde{C} \cdot \tilde{\beta} + \Xi, \tag{3}$$

where $\tilde{X}$ is an np column vector, $\tilde{C}$ is an $(np)\times[(d+1)p]$ block diagonal matrix and $\tilde{\beta}$ is a $(d+1)p$ column vector that contains all gene-specific information. The full and the reduced model can both be written this way. Although the model is multivariate, computation is quite feasible, because the blocks of $\tilde{C}$ are all identical design matrices for each gene-wise model. The noise component has a mean of 0 and an unspecified positive definite covariance matrix $\mathrm{Cov}[\Xi] = \tilde{\Sigma}$. Based on the global linear model, it is possible to estimate the residuals $\tilde{R}$ and the residual sum of squares (RSS) based on the hat matrix $\tilde{H}$:

$$\begin{aligned} \tilde{R} &= (\Delta_{np} - \tilde{C}(\tilde{C}^t\tilde{C})^{-1}\tilde{C}^t)\tilde{X} = (\Delta_{np} - \tilde{H})\tilde{X} \\ \text{and} \quad \mathrm{RSS} &= \tilde{R}^t \cdot \tilde{R} = \tilde{X}^t(\Delta_{np} - \tilde{H})\tilde{X}. \end{aligned} \tag{4}$$

Here Δ_{np} is the np dimensional identity matrix. The global RSS can be computed easily as the sum of gene-wise residual sums of squares. The extra sum-of-squares principle is used to study the difference between the full model and the reduced model (Draper and Smith 1998). It allows the construction and computation of a multivariate test statistic:

$$F_{GA} = \frac{\mathrm{RSS_{RM}} - \mathrm{RSS_{FM}}}{\mathrm{RSS_{FM}}} \cdot \frac{n-q}{f}, \tag{5}$$

where n is the number of samples, f is the number of columns in C_0 (namely, the number of parameters in the full model minus the number of parameters in the reduced model) and q is the number of columns in $[C_0, C_1]$ (number of parameters in the full model).

Under the assumption of independent homoscedastic gene expressions $\Xi \sim N(0, \sigma^2\Delta_{np})$ the null hypothesis that there is no group influence on global gene expression $H_0 : \tilde{\beta}_0 = 0$ can be tested using a classical F-test. The test statistic F_{GA} is $F_{pf,p(n-q)}$ distributed under H_0. In general, the simple assumption of independent homoscedastic genes does not hold.

Table 2 specifies the full and reduced model for six specific and often used scenarios.

2.2 Permutation distribution of F_{GA} under the reduced model

The implementation of a permutation-based approximation for the distribution of F_{GA} under the null hypothesis is straightforward. One permutes the rows of C_0 B times, which corresponds to subject sampling. Because rows of C_1 are not permuted the covariate structure is preserved. To reduce the computational burden, we take the residuals of the reduced model from the original data as fixed and fit these residuals to the permuted design matrix $[C_0^b, C_1]$ in order to calculate the residuals for the full model. Therefore, the $\mathrm{RSS_{RM}}$ is constant for each permutation. Only $\mathrm{RSS_{FM}}$ varies and comes up with the resampled value of the statistic F^b_{GA}. An empirical P-value is given by the fraction of statistics F^b_{GA} $(b = 1, \ldots, B)$ that are larger than the actual F_{GA}.

2.3 Asymptotic distribution of F_{GA} under the reduced model

The expressions for the p genes in a sample are assumed to be normally distributed with an unspecified covariance matrix $\tilde{\Sigma}_{\mathrm{genes}}$. The null hypothesis states that the mean of the gene expression is determined by the reduced model. Basic algebra proves that the difference between both RSSes is distributed as a mixture of χ^2 distributions:

$$\mathrm{RSS_{effect}} = \mathrm{RSS_{RM}} - \mathrm{RSS_{FM}} = \sum_{i=1}^{np} \pi_i \cdot \chi^2_{1,i}. \tag{6}$$

Kotz *et al*. (1967) describe an algorithm to approximate the distribution of the weighted sum of χ^2-distributed variables by a (possibly infinite) mixture of χ^2 distributions.

The calculation of the np weights needs two inputs: the n eigenvalues $\rho_1, \ldots, \rho_n$ (possibly the same numeric value with a corresponding multiplicity) of the difference of the hat matrices $(H_{\mathrm{FM}} - H_{\mathrm{RM}})$ and the p eigenvalues $\lambda_1, \ldots, \lambda_p$ of the gene covariance matrix $\tilde{\Sigma}_{\mathrm{genes}}$. While the ρ's can be calculated by standard methods, the calculation of the λ's is not straightforward. A regularized estimate of $\tilde{\Sigma}_{\mathrm{genes}}$ is needed in case the number of genes is larger than the number of samples. This implies an empirical covariance matrix not of full rank. Ledoit and Wolf (2004)

Table 2. Six specific and often used scenarios in model notation of the S language

Design	Full model $[C_0, C_1]$	Reduced model $[C_1]$
Groups	$\sim group + cov$	$\sim cov$
Dose- response	$\sim dose + cov$	$\sim cov$
Group by dose interaction	$\sim group * dose + cov$	$\sim group + dose + cov$
Time trends in groups	$\sim group * time + cov$	$\sim group + time + cov$
Gene gene interaction (GGI)	$\sim gene + cov$	$\sim cov$
Differential GGI	$\sim group * gene + cov$	$\sim group + gene + cov$

propose an estimate $\Sigma_\varphi = \varphi \cdot T + (1-\varphi) \cdot U$ with shrinkage factor φ, shrinkage target T and unrestricted estimate U. The optimal shrinkage factor φ can be explicitly computed from the data for a given full rank positive definite shrinkage target T. If few genes are correlated, the shrinkage target may be the diagonal matrix with unequal variances. For this target the optimal φ is

$$\varphi^* = \frac{\sum_{i \neq j} \mathrm{Var}(s_{ij})}{\sum_{i \neq j} s_{ij}} \quad (7)$$

where s_{ij} is an unbiased estimate of the covariance between gene i and gene j. The calculation can be performed by the function `cov.shrink` of the R-package `corpcor` (Schaefer *et al.* 2006).
The np weights π are:

$$\pi = \{\lambda_i \cdot \rho_j,\ i = 1, \ldots, p;\ j = 1, \ldots, n\}.$$

3 SIMULATION STUDY

The simulation study assesses the quality of the permutation-based P-value and the asymptotic P-value. The calculation of the asymptotic P-value involves two crucial approximation steps: a shrinkage approach to estimate the high-dimensional covariance matrix and a finite approximation to the series expansion of the distribution function for the weighted χ^2 distribution. Both steps may influence the performance of the test.

The permutation P-value will be compared with the asymptotic P-value. In the case of independent genes, the statistical theory provides a theoretical P-value derived from an F-distribution. Under the null hypothesis, the P-value should be uniformly distributed on [0,1]. The simulation setup can also be used to study the power of the test.

Four scenarios will be studied following a common scheme: one thousand permutations are used to calculate the permutation P-value, and one thousand independent repeats of the experiment are performed.

The first scenario (S1) studies independent genes with no differential expression between two groups of samples: 30 (200) independent $N(0,1)$ distributed genes with 20 samples taken in each group.

The second scenario (S2) looks at 30 (200) dependent $N(0,1)$ distributed genes with 20 samples taken in each group. Dependence is defined by compound symmetry: an equally positive correlation between two genes ($\rho = 0.2$). This scenario is a challenge to the shrinkage target used in the algorithm. The default shrinkage target is the diagonal matrix while the ideal shrinkage target for S2 would be a matrix with a common covariance. This dependence structure represents regulatory networks with genes ordered in a chain and partial correlation between two neighbouring genes.

The third scenario (S3) is based on the second scenario. Additionally 6 (20) genes are randomly chosen to be differentially expressed with a mean difference of 0.5 between the groups.

The fourth scenario (S4) looks at networks that consist of short chains of interacting genes. This can be represented by small blocks of compound symmetry within the large covariance matrix of the gene group. For simulation purposes we chose blocks of size 10. We look at the null hypothesis and at the alternative as described in S3.

Table 3. False positive fraction (top) and power (bottom) at $\alpha = 5\%$

Scenario	Level permutation	Level asymptotic
S1-30genes	0.055	0.062
S1-200genes	0.046	0.069
S2-30genes	0.046	0.059
S2-200genes	0.057	0.101
S4-30genes	0.049	0.073
S4-200genes	0.049	0.070
Scenario	**Power permutation**	**Power asymptotic**
S3-30genes	0.204	0.283
S3-200genes	0.121	0.225
S4-30genes	0.366	0.425
S4-200genes	0.647	0.732

The results of the simulation study are summarized in Table 3. The simulation for the permutation P-values shows a behaviour according to the uniform distribution on [0,1] in the four settings of S1 and S2. S3 is under the alternative and a strong deviation from the uniform distribution is expected. The accurateness of the permutation P-values is in the expected range around the theoretical P-value (10%, 90% Quantile ~2%) and can be improved by increasing the number of permutations.

The simulation for the asymptotic P-values shows an anti-conservative behaviour that produces more false positive signals than expected by the level of the test. The observed levels for the nominal 5% are 6% (S1–30genes), 7% (S1–200genes), 6% (S2–30genes) and 10% (S2–200genes). The anti-conservative behaviour has an impact on the power of the asymptotic test. Under S3, the asymptotic (permutation) test on level 5% shows a power of 28% (20%) for 30 genes and 23% (12%) for 200 genes.

The asymptotic test improves under S4. Its distribution is as expected similar to the uniform. The observed levels for the nominal 5% are 7% (S1–30genes) and 7% (S1–200genes). Under S4, and the alternative described in S3 the asymptotic (permutation) test on level 5% shows a power of 43% (37%) for 30 genes and 73% (65%) for 200 genes.

We conclude that the permutation P-value works sufficiently accurately but is extremely computationally demanding when exploring low P-values in a multiple testing setting. The asymptotic P-value can replace the permutation approach in situations of complex multiple testing. In general, it is slightly anti-conservative when the shrinkage target is chosen appropriately. A misspecification of the shrinkage target (for instance, choosing the diagonal matrix when compound symmetry is present) may amplify the anti-conservative tendency and result in an uncontrolled rate of false significant findings.

4 GRAPHICAL DISPLAY OF TEST RESULTS

As mentioned above, the GlobalANCOVA procedure is based on the extra-sum-of-squares principle (Draper and

Smith, 1998). Therefore, the decomposition of the total sum of squares with respect to model components can be studied in two ways: as the sum of contributions per gene or as the sum over the subject contributions. This allows a gene-wise and a subject-wise view and visualizes which genes or which samples are most affected in their gene expression by the structure under study. We have adapted these views into two graphical displays of the results.

The function `Plot.genes` displays the gene-wise reduction of the sum of squares

$$\mathrm{RSS}_{\mathrm{gene}i} = \sum_{j=1}^{n} \hat{\xi}^2_{\mathrm{RM},ij} - \hat{\xi}^2_{\mathrm{FM},ij}$$

divided by the difference in degrees of freedom between the full and the reduced models. Gene-wise values are shown as bars. A reference line corresponds to the gene-wise residual mean-square of the full model. This plot can be regarded as a representation of gene-wise F-tests of the single-gene hypotheses. It shows the genes that contribute most to the structural differences in expression looked for.

In addition, a subject-wise view is provided by the function `Plot.subjects`. It represents a comparison of the fit between the full and the reduced models for each sample:

$$\mathrm{RSS}_{\mathrm{sample}\ j} = \sum_{i=1}^{p} \hat{\xi}^2_{\mathrm{RM},ij} - \hat{\xi}^2_{\mathrm{FM},ij}$$

The sum over all gene-wise contributions per sample indicates the improvement in fit per sample over all genes. The values need not be positive. Small or negative values indicate that the fit for a given sample is not improved by including the structural terms of interest.

Both plots can be coloured with respect to a design variable of interest. The meaning of colouring is straightforward in the subjects plot because each subject has a unique value for the variable under consideration. The colouring of the gene plot is driven by the mean value of gene expression in the subgroups defined by the values of the variables of interest. A gene is coloured with respect to the subgroup where its mean gene expression is highest. Examples of both plotting types are given in the example section.

5 EXAMPLES

5.1 Two groups with differential gene signatures

Groene *et al*. (2006) studied the p53 pathway, in order to differentiate colorectal carcinoma (CRC) patients in UICC stage II (good prognosis) versus UICC stage III (bad prognosis). While UICC II CRC has a 5-year recurrence rate of 20–25%, UICC stage III tumours are more dynamic with a 5-year recurrence rate over 40% after radical resection (Obrand and Gordon 1997).

Tumour samples of 18 patients with UICC stage II CRC and 18 patients with UICC stage III CRC were hybridized on U133A Affymetrix GeneChips. Forty-five probesets of the U133A Affymetrix GeneChip were associated with the p53-signalling pathway. A simple initial analysis may consist in studying the unadjusted influence of group (UICC II/UICC III) on the gene expression for the ensemble of the 45 probesets of interest. In this case, the full and the reduced models are defined by

$$C_{\mathrm{FM}}^{\mathrm{unadjusted}} = \begin{pmatrix} 1 & g_1 \\ \vdots & \vdots \\ 1 & g_{36} \end{pmatrix} \quad \text{and} \quad C_{\mathrm{RM}}^{\mathrm{unadjusted}} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

where $g_i = 0$ if sample i is from a UICC II patient, $g_i = 1$ if sample i is from a UICC III patient. The value of the test statistic F_{GA} is 2.2469. The permutation P-value derived from 10 000 resamples is $p_{\mathrm{perm}} = 0.011$.

Colorectal carcinomas are located in two parts of the gut, the colon and the rectum. A carcinoma of the rectum has slightly different biological properties than the carcinoma of the colon. Thus, it may be reasonable to consider the location of the tumour when modelling its gene expression. Unlike breast cancer, which is rarely found in men, CRC is present in both sexes. Sex also influences gene expression and one should adjust for it when both groups are not homogeneous with respect to sex. We perform a gene-wise adjustment to both covariates (location: $l_i = 0$ if the CRC of patient i is located in the colon, $l_i = 1$ else; sex: $s_i = 1$ if patient i is male, $s_i = 0$ else). In the adjusted case, the full and the reduced model are defined by

$$C_{\mathrm{FM}}^{\mathrm{adj}} = \begin{pmatrix} 1 & g_1 & l_1 & s_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & g_{36} & l_{36} & s_{36} \end{pmatrix} \quad \text{and} \quad C_{\mathrm{RM}}^{\mathrm{adj}} = \begin{pmatrix} 1 & l_1 & s_1 \\ \vdots & \vdots & \vdots \\ 1 & l_{36} & s_{36} \end{pmatrix}$$

where g_i is defined as above. The value of the test statistic F_{GA} is 2.8099. The permutation P-value derived from 10 000 resamples is $p_{\mathrm{perm}} = 0.002$.

The asymptotic distribution of F_{GA} under the null hypothesis that grouping has no influence on the gene signature of a gene group can now be derived for the unadjusted case ($\mathrm{RSS}_{\mathrm{effect}} = 14.724$) and the adjusted case ($\mathrm{RSS}_{\mathrm{effect}} = 18.079$) [see formula (6)]. The eigenvalues $\{\lambda_i\}_{i=1,\ldots,45}$ of the covariance matrix for the multivariate distribution of the gene expression can be calculated using the algorithms provided by Schaefer *et al*. (2006). The 36 eigenvalues of the 36 by 36 matrix $H_{FM} - H_{RM}$ in the unadjusted and the adjusted situation are 1 and 35 times 0. The resulting mixture of χ^2 distributions consists of 45 components with weights $\{\lambda_i\}_{i=1,\ldots,45}$. The algorithm described above gives the asymptotic P-values of 0.0102 (unadjusted) and 0.0018 (adjusted) that agree quite well with the permutation P-values.

The gene plot (Fig. 1, top) shows two strong grey signals for probesets that are upregulated in the group of patients with advanced cancer (UICC III). The two related genes, CDKN2A and CDKN1C, are two cyclin-dependent kinase inhibitors, whose association with colorectal carcinogenesis has been previously shown (Li *et al.,* 2003; Maeda *et al*., 2003). The sample plot (Fig. 1, bottom) shows only a few black bars into the negative direction: the full model fits the gene expression of patients in the good prognostic group (UICC II, black bars) better. The gene expression of the p53 pathway is less predictable in the group of patients with advanced cancer (UICC III, grey bars).

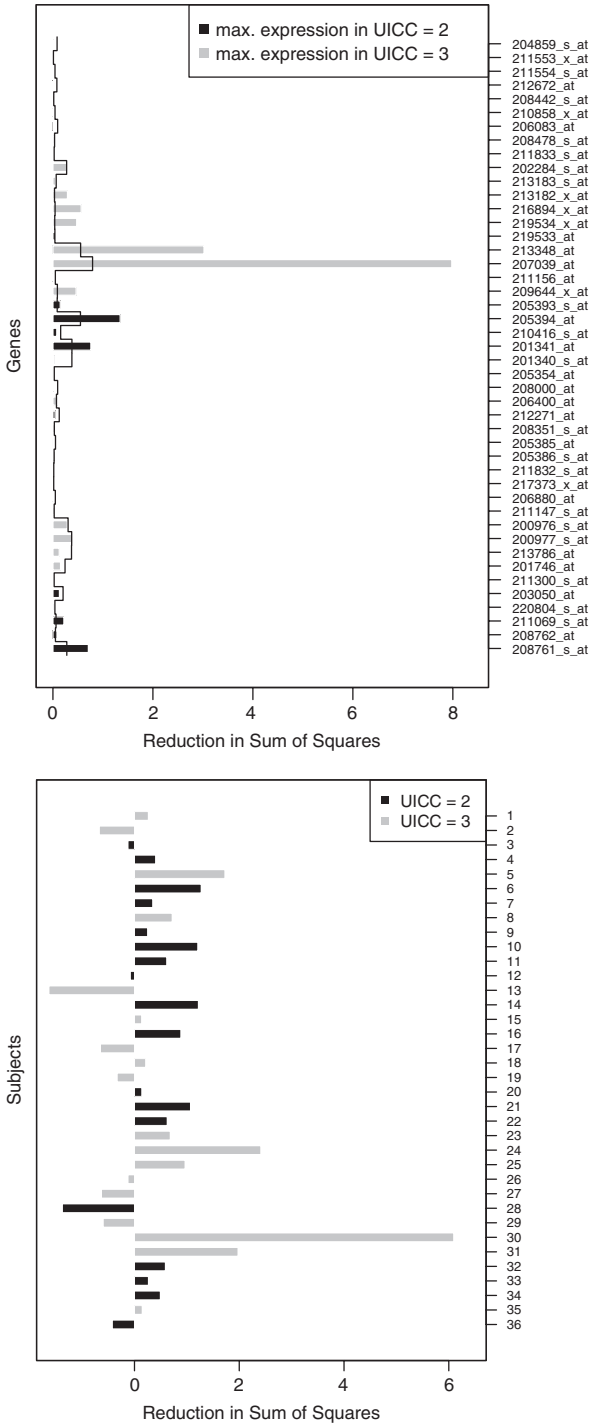


Fig. 1. Gene (top, labels = probeset ID) and subject (bottom, labels = patient ID) plot for the p53 example.

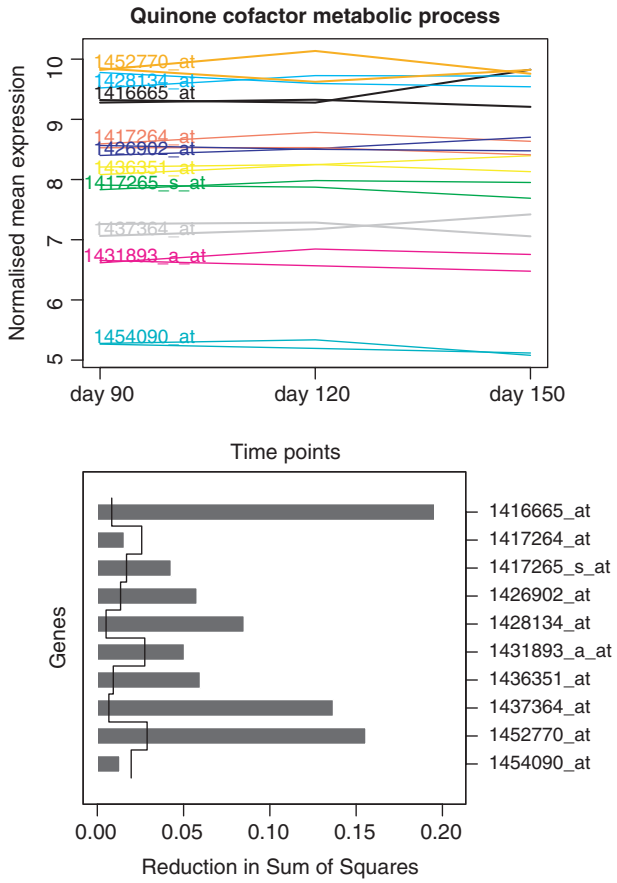
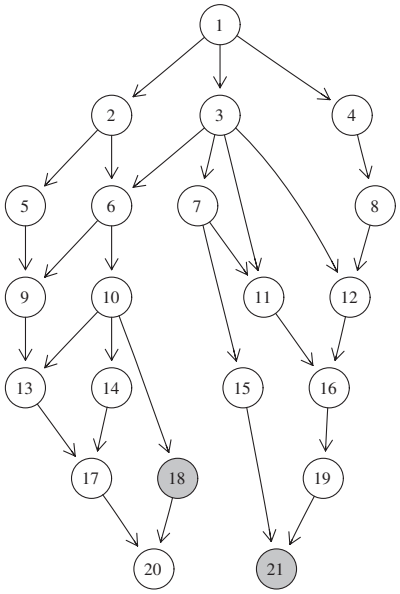


Fig. 2. Top: time course of expression for genes in biological process GO category 'quinone cofactor metabolic process', which was found to be interesting with respect to different temporal development in expression between the two treatment groups. Each gene is coloured uniquely. The two lines for each gene correspond to normalized mean expression values in the prion-infected and mock-infected groups at each time point. Genes with visible interaction between treatment group and time are shown with thick lines. Bottom: contribution of single genes in the biological process GO category 'quinone cofactor metabolic process' to the GlobalANCOVA test statistic.

5.2 Differential time course

Xiang *et al*. (2007) compared gene regulation over time in a group of mice infected by prion ME7 compared to a group of mice with mock-infected brains (inoculated with normal brain homogenate). Three mice were sacrificed in each group at each of three time points (90/120/150 days after infection). The genes of interest to study were those that are differentially regulated over time between both groups. Time was modelled as an ordered factor in a linear model and polynomial contrasts were used. The main interest of the experiment was the relevance of the time by group interaction to explain the observed data (implying a differential time course between the groups).

The probesets of the moe430a GeneChip were mapped to the Gene Ontology (Gene Ontology 2000). For each node in the biological process (BP) ontology, a GlobalANCOVA was performed to compare the full model ($\sim group * time$) with the reduced model ($\sim group + time$). Mining all GO nodes of the BP ontology may detect gene groups that might play an important role in the differentiation of gene expression over time between the two experimental groups. The term with the smallest P-value ($p = 0.00003$) is *quinone cofactor metabolic process*. The time course and gene plot are shown in Figure 2.



	ID	Term		ID	Term
1	8150	biological process	12	50794	regulation of cellular proc.
2	8152	metabolic process	13	51188	cofactor biosynthetic proc.
3	9987	cellular process	14	6732	coenzyme metabolic process
4	65007	biological regulation	15	278	mitotic cell cycle
5	9058	biosynthetic process	16	9108	coenzyme biosynthetic proc.
6	44237	cellular metabolic proc.	17	9108	coenzyme biosynthetic proc.
7	7049	cell cycle	18	42375	quinone cofactor metabol. proc.
8	50789	regulation of biol. proc.	19	74	regul. of progr. thr. cell cycle
9	44249	cellular biosynth. proc.	20	45426	quinone cofactor biosynth. proc.
10	51186	cofactor metabolic proc.	21	7346	regul. of progr. thr. mitotic c.c.
11	22402	cell cycle process			

Fig. 3. Focus level graph for the BP ontology showing gene groups with a differential time course in gene regulation between mock-infected and scrapie-infected mice. Strong FWER is controlled on level $\alpha = 0.1$. Gene sets within the chosen focus level are coloured grey.

Genes showing a clear differential time course of expression in the upper figure are also detected by the gene plot to have the most influence on the GlobalANCOVA test statistic.

Some adjustment for multiple testing is needed because several thousands of GO categories are tested. The Bonferroni–Holm correction on a global level of $\alpha = 0.1$ does not return any node as significant. The *focus level procedure* of Goeman and Mansmann (2007) offers a more efficient alternative for multiple testing on the GO graph (strong control of the family-wise error rate). It combines the strengths of *closed testing* with *Bonferroni–Holm* and starts the search for relevant nodes in a middle section of the GO graph. The method is available in the Bioconductor (http://www.bioconductor.org) packages `globaltest` and `GlobalAncova`. The *focus level procedure* determines a subgraph of the GO with a controlled number of falsely rejected null hypotheses (*no gene in the specific BP subgroup shows differential time course*). For the data at hand, the procedure returns the Figure 3 on a global significance level of $\alpha = 0.1$. The graph contains 3 of the 5 BP groups that are detected by the FDR controlling procedure of Benjamini–Hochberg on a level of 10%. The two additional nodes are *telomere organization and biogenesis* (GO:0032200) and *telomere maintenance* (GO:0032200). Because the experiment only involved 18 animals, the global level of testing was set to be more liberal on $\alpha = 0.1$.

6 DISCUSSION

GlobalANCOVA is a general methodology for analysing gene expression data in terms of predefined gene sets, pathways or complexes. Its constructive idea is to use gene-wise linear models and to aggregate their information in a multivariate test procedure. GlobalANCOVA exploits the strength of the classical linear model theory, especially ideas related to goodness of fit tests. In this article, we extend the GlobalANCOVA proposed by Mansmann and Meister (2005) to a general framework that makes full use of the refined theory for linear models. In the previous version merely the typical question about two-class differential expression could be addressed, whereas now the approach is broadened to a wide field of applications. Basically, it allows for testing whether a specific aspect of the study design such as group membership, time course, group by time course interaction, dosage, group by dose interaction, etc. is necessary to explain the observed gene expression. Another important methodological advancement makes asymptotic *P*-values available that helps to speed up the calculation.

The usefulness of linear models for the analysis of high-dimensional experiments is widely acknowledged and exploited. GlobalANCOVA generalizes approaches as used in limma (Smyth, 2005) from experimental design to observational studies by allowing the inclusion of covariates that may correct for unbalanced situations. GlobalANCOVA uses the full strength of linear model theory: model building techniques, strategies for model diagnostics and model selection. Other proposals for global tests (Goeman *et al.*, 2004; Kong *et al.*, 2006; Tomfohr *et al.*, 2005) restrict their applicability to specific designs (groups, main effects).

Goeman and Bühlman (2007) differentiate *competitive* versus *self-contained* tests for gene groups. A competitive test compares differential expression of the gene set to a standard defined by the complement of that gene set. A self-contained test, in contrast, compares the gene set to a fixed standard that does not depend on the measurements of genes outside the gene set. Goeman and Bühlman (2007) give a thorough discussion on the pros and cons of both concepts. GlobalANCOVA is a self-contained test for gene groups.

The self-contained test can be performed from two perspectives: prediction and structure. The global test proposed by Goeman *et al.* (2004) assesses the predictive power within gene expression data X for a certain phenotype Y. Here, phenotypic covariates C and gene expression are on the same footing. Its null hypothesis is that the knowledge of gene expression does not improve the prediction for Y: $P[Y \mid X,C] = P[Y \mid C]$. GlobalANCOVA studies the effect of a design on gene expression patterns: $P[X \mid Y,C] = P[X \mid C]$. Does the time course ($Y$ represents the variables that encode the time by group interaction) of gene expression depend on group membership (C encodes the main factors `time` and `group`)? The differential time course design could not be treated by `globaltest`.

Global tests are designed to reject a very general null hypothesis: no gene in a group of genes shows a reaction of interest. This null hypothesis offers an umbrella for two extreme cases: there are few strongly reacting genes or there are many weakly reacting genes. Furthermore, global tests provide a first step for a more refined analysis. They offer a proof of concept before searching the specific aspects of a cellular mechanism. The global test of Goeman *et al*. (2004) can be seen as a first check before undertaking the complex task of building a classification rule. Global tests can be used to validate statements regarding gene sets.

In spite of their quite general view, global tests can be used to build refined pictures on a phenomenon under study. The specification of a question can be represented by a hierarchy of null hypotheses. Statistical strategies for multiple testing can be used to control the error when the question of interest is framed by a global test. Goeman and Mansmann (2007) describe the focus-level method as a tool to locate substructures in a GO graph where gene expression is related to a specific biological occurrence. The specification of a question may be encoded in the most specific nodes of the detected GO substructure. Using global tests helps to control the family wise error rate of the derived statement. Meinshausen (2007) developed an alternative procedure to derive sound statements from tree-structured hierarchies of null hypotheses. Methodologies to control hierarchical false discovery rates are proposed by Yekutieli (2007).

The category approach by Gentleman and Falcon (2007) is similar in spirit to GlobalANCOVA. The gene-wise application of the full model returns a set of regression coefficients and the summary statistic is their mean. The summary statistic should be approximately normally distributed under the null hypothesis (given by the reduced model) with mean zero. The category package provides visual tools to check this assumption for one or a few groups. A permutation procedure similar to the procedure described in subsection 2.2 can be used to derive the distribution of the summary statistic under the null hypothesis and to compare it with the observed value. There is no proposal to derive an approximate *P*-value, as is necessary when looking for minuscule *P*-values or when performing multiple testing on the GO.

In summary, global tests offer an essential tool within strategies to mine high-dimensional data based on structured biological knowledge. The usefulness of group testing in terms of stable and reproducible findings on gene expression has also been confirmed by other groups (Manoli *et al*., 2006).

From a purely statistical point of view, GlobalANCOVA is just a replacement for the general multivariate linear model analysis in very high dimensions. Its strength for the application in gene expression analysis lies in its flexibility to incorporate substantial biological information via measured covariates and in its ability to model complex phenotype-related effects.

ACKNOWLEDGEMENT

This work was supported by the NGFN project 01 GR 0459, BMBF, Germany.

Conflict of Interest: none declared.

REFERENCES

Draper,N.R. and Smith,H. (1998) *Applied Regression Analysis*. 3rd edn. Wiley-Interscience, New York.

The Gene Ontology ConsortiumGene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Gentleman,R. and Falcon,S. (2007) Category: Category Analysis. R package version 2.1.30.

Goeman,J.J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Goeman,J.J. and Bühlmann,P. (2007) Methodological issues in gene set testing based on microarray data. *Bioinformatics*, **23**, 980–987.

Goeman,J.J. and Mansmann,U. (2007) Multiple testing on the directed acyclic graph of gene ontology. *Technical report*. http://www.msbi.nl/dnn/People/Goeman/Publications/tabid/202/Default.aspx.

Groene,J. *et al*. (2006) Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III. *Int. J. Cancer*, **119**, 1829–1836.

Kerr,M.K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol*., **7**, 819–837.

Kong,S.W. *et al*. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Kotz,S. *et al*. (1967) Series representations of distributions of quadratic forms in normal variables. I. Central case. *Ann. Math. Stat*., **38**, 823–837.

Lamb,J. *et al*. (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*, **114**, 323–334.

Ledoit,O. and Wolf,M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal*., **88**, 365–411.

Li,J.Q. *et al*. (2003) Loss of p57KIP2 is associated with colorectal carcinogenesis. *Int. J. Oncol*., **23**, 1537–1543.

Maeda,K. *et al*. (2003) Hypermethylation of the CDKN2A gene in colorectal cancer is associated with shorter survival. *Oncol. Rep*., **10**, 935–938.

Manoli,T. *et al*. (2006) Group testing for pathway analysis improves comparability of different microarray data sets. *Bioinformatics*, **22**, 2500–2506.

Mansmann,U. and Meister,R. (2005) Testing differential gene expression in functional groups. *Methods Inf. Med*., **44**, 449–453.

Meinshausen,N. (2007) Hierarchical testing of variable importance. *Technical report*. http://www.stats.ox.ac.uk/~meinshau/hierarchical.pdf.

Obrand,D.I. and Gordon,P.H. (1997) Incidence and patterns of recurrence following curative resection for colorectal carcinoma. *Dis. Colon Rectum*, **40**, 15–24.

Schaefer,J. *et al*. (2006) corpcor: Efficient Estimation of Covariance and (Partial) Correlation. R package version 1.4.4. http://www.strimmerlab.org/software/corpcor/.

Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R. *et al*. (eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Tomfohr,J. *et al*. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.

Xiang,W. *et al*. (2007) Transcriptome analysis reveals altered cholesterol metabolism during the neurodegeneration in mouse scrapie model. *J. Neurochem*, **102**, 834–847.

Yekutieli,D. (2007) Hierarchical False Discovery Rate controlling methodology. Accepted by the Journal of the American Statistical Association.